Equivalentie

van de Vijver, F.J.R.

Published in:
Computers in Psychologie

Publication date:
1994

Equivalentie: Terugblik en Vooruitblik

F. van de Vijver

Samenvatting

Empirisch onderzoek toont meestal aan dat de computer-ondersteunde en papier-en-potlood afnames van persoonlijkheidsvragenlijsten equivalent zijn. Bij capaciteitentests is dit slechts het geval als de taak zonder tijdlimiet afgenomen wordt. Resultaten op capaciteitentests met een snelheidslimiet kunnen sterk beïnvloed worden door de afnamevorm. Met name eenvoudige tests waarbij snel en nauwkeurig gewerkt moet worden, kunnen een beroep doen op wat andere vaardigheden in beide testafnames. Er wordt gepleit voor een gedetailleerde analyse van de psychologische verschillen in beide testvormen. Een empirisch onderzoek naar de equivalentie van de conventionele en computer-ondersteunde versie van de General Aptitude Test Battery wordt kort beschreven.

Inleiding

Gedurende de laatste jaren hebben we een groot aantal studies gezien naar de equivalentie van computer-ondersteunde en papier-en-potlood tests (hier verder te noemen conventionele tests) (zie o.a. Greaud & Green, 1986; Mead & Drasgow, 1993; Neubauer, Urban, & Malle, 1991). In het eerste deel van dit artikel zal een kort overzicht gegeven worden van de recente stand van zaken op het gebied van equivalentie-onderzoek. In het tweede deel zal ingegaan worden op een onderzoek naar de equivalentie van twee afnamevormen van de General Aptitude Test Battery (GATB; Van de Vijver & Harsveld, in druk). In het derde deel zal ingegaan worden op de vraagstellingen die in mijn visie nog verder onderzoek behoeven.

Er zal betoogd worden dat de grote lijnen uit het equivalentie-onderzoek intussen wel duidelijk zijn (cf. Akkerman, 1994); bij persoonlijkheidstests heeft de introductie van de computer meestal geen noemenswaardige gevolgen als maar niet met de introductie van de computer ook allerlei andere aspecten van de test gewijzigd zijn. Te denken valt onder andere aan de anonieme afname via de computer en de meer persoonlijke afname van conventionele persoonlijkheidsvragenlijsten. Bij capaciteitentests dient een onderscheid gemaakt te worden in tests met en tests zonder tijdlimiet. Instrumenten met een sterk snelheidskarakter laten vaak andere resultaten zien bij een computer-ondersteunde afname dan bij een conventionele afname. Resultaten op tests zonder tijdlimiet worden meestal niet of nauwelijks beïnvloed door het afnamemedium (Mead & Drasgow, 1993).

Er zijn al veel equivalentiestudies verricht waarin een test op twee manieren afgenomen wordt. Deze studies zijn nodig om inzicht te krijgen in de equivalentie van beide testafnames, maar het valt niet te verwachten dat nieuwe studies onze theoretische inzichten veel zullen vergroten. Vanuit theoretisch perspectief lijken de dagen van dergelijk equivalentie-onderzoek geteld. Toch blijven er nog intrigerende vragen over. Deze hebben met name betrekking op een analyse van de verschillen tussen beide testafnames: Waarom verandert de psychologische betekenis van snelheidstests soms zo ingrijpend als deze middels een computer afgenomen worden? Verder is nog onvoldoende bekend over mogelijke verschillen in predictieve validiteit van beide soorten afnames. Er is weinig reden om te veronderstellen dat er grote verschillen zullen optreden in predictieve validiteit tussen computer-ondersteunde en conventionele testafnames, met name als de tests equivalent zijn, maar het is nog niet aangetoond.

Equivalentie: de belangrijkste bevindingen

De American Psychological Association heeft de volgende definitie gegeven van score-equivalentie:

Scores from conventional and computer administrations may be considered equivalent when (a) the rank orders of scores of individuals tested in alternative modes closely approximate each other, and (b) the means, dispersions and shapes of the score distributions are approximately the same, or have been made approximately the same by rescaling the scores from the computer mode (geciteerd in Green, 1991, p. 248).

In de literatuur zijn veel onderzoekingen gerapporteerd naar de equivalentie van computer-ondersteunde en conventionele testafnames (een overzicht is te vinden in Mead & Drasgow, 1993). Met name Raven's Matrijzen lijken onderzoekers telkens weer te inspireren (bijv. Neubauer, Urban, & Malle, 1991). De uitkomsten van deze onderzoekingen lopen vaak behoorlijk uiteen. Voor een deel zal dit te maken hebben met het gebrek aan standaardisatie van hardware en software. De meeste bekende computermerken, operating systems, soorten monitoren, schermresoluties en response devices zijn wel gebruikt in het onderzoek. Voor een deel kan het ook te maken hebben met de populatie waarbij de test afgenomen is. De populaties betreffen onder andere psychogeriatrische en psychiatrische patiënten, schoolkinderen en studenten.

Gegeven deze variëteit is het des te opmerkelijker dat Mead en Drasgow (1993) in een meta-analyse van studies naar de equivalentie van cognitieve tests tot zo'n duidelijke conclusie konden komen. De auteurs vonden een voor meetonbetrouwbaarheid gecorrigeerde correlatie tussen computer-ondersteunde en conventionele afnames van .97 voor tests zonder (of zonder een noemens-

waardige) tijdlimiet en van .72 voor tests met een tijdlimiet. De introductie van de computer lijkt dus met name gevolgen te hebben voor snelheidstests. Verder vonden de auteurs een verwaarloosbaar verschil in gemiddelde scores tussen beide testafnames: .03 standaarddeviatie voor tests zonder tijdlimiet en .07 standaarddeviatie voor tests met een tijdlimiet. Mead en Drasgow rapporteren geen gegevens over eventuele verschillen in betrouwbaarheid, maar de conclusie dat computer-ondersteunde testafnames vrijwel nooit een lagere betrouwbaarheid te zien geven lijkt gerechtvaardigd op basis van de literatuur.

Een voordeel van het gebruik van computers in testafnames is de beschikbaarheid van behoorlijk wat rekenkracht tijdens de testafname. Adaptieve tests maken daar goed gebruik van. Het afnemen van te gemakkelijke items aan vaardige proefpersonen of te moeilijke items aan weinig vaardige proefpersonen waardoor de motivatie kan verdwijnen, kan tegengegaan worden door het gebruik van adaptieve tests. Mead en Drasgow vonden dat het gebruik van adaptieve tests geen invloed heeft op het gemeten construct. Wel dient toegevoegd te worden dat deze bevindingen op slechts enkele onderzoekingen berusten en dat de wellicht meest populaire slotzin van wetenschappelijke rapporten "more research is needed" hier niet alleen een cliché is.

Er is geen meta-analyse uitgevoerd op studies naar de equivalentie van beide manieren van afname van persoonlijkheidsvragenlijsten. De reden hiervoor lijkt me dat de mogelijke invloed van de methode van afname op de scores sowieso al lager ingeschat wordt. Het lijkt redelijk om equivalentie te verwachten als tenminste de overige testomstandigheden niet teveel verschillen in beide condities. Zoals Akkerman (1994) al aangaf, kunnen overwegingen van privacy een rol gaan spelen. Een anonieme computer en een veel persoonlijker testafname door een psycholoog zouden wel eens andere responsen op kunnen roepen vooral als het om persoonsgevoelige informatie gaat.

Nu duidelijk is welke kant de invloed van de introductie van computers bij testafnames opgaat, blijven nog andere vragen over. Deze hebben met name betrekking op verschillen tussen beide media in het geval van snelheidstests. Welke psychologische verschillen treden op tussen conventionele en computerondersteunde afnames van snelheidstests? In eigen onderzoek van Van de Vijver en Harsveld (in druk) is geprobeerd deze vraag op te lossen voor de GATB.

De onvolledige equivalentie van de GATB

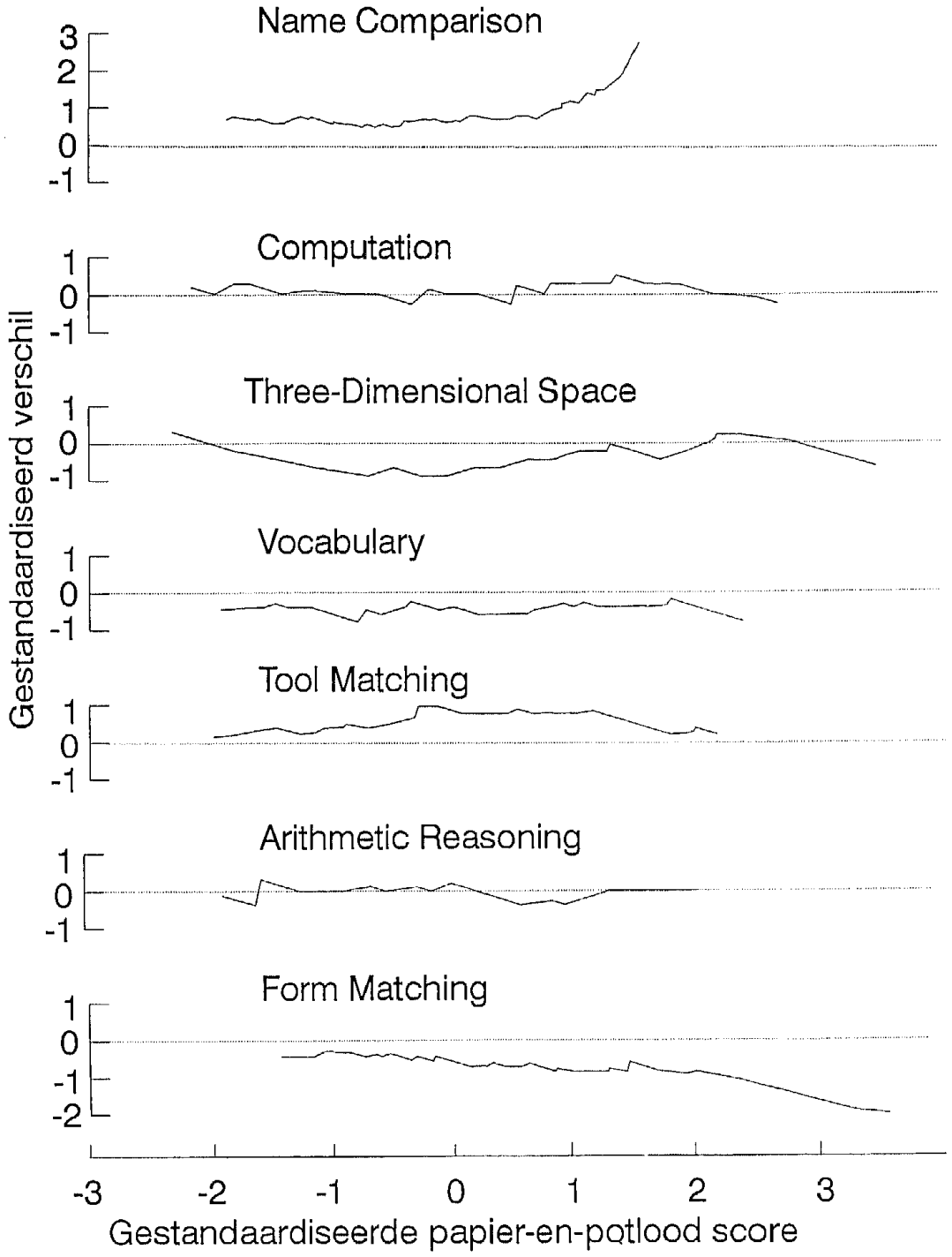
De conventionele en de computer-ondersteunde versie van de GATB zijn elk afgenomen bij 163 sollicitanten voor de Koninklijke Militaire Academie. De groepen waren samengesteld via een matching op leeftijd, geslacht en intelligentie (gemeten met de Berenschot Intelligentietest). De GATB is een algemene

intelligentietest met een sterke nadruk op snelheid van reageren. Er zijn zeven subtests: *Name Comparison* (twee kolommen met namen worden gepresenteerd; de proefpersoon moet aangeven of de namen gelijk of verschillend zijn), *Computation* (optellen, aftrekken, vermenigvuldigen en delen van gehele getallen), *Three-Dimensional Space* (er is een drie-dimensionele "targetfiguur" en vier alternatieve figuren; proefpersoon moet aangeven welke van de vier door rotatie identiek aan de targetfiguur kan worden), *Vocabulary* (in een groep van zes woorden bevinden zich twee identieke of tegengestelde woorden), *Tool Matching* (welke van de vier figuren is identiek aan een "target"?), *Arithmetic Reasoning* (redactiesommen) en *Form Matching* (twee sets figuren worden gepresenteerd. De proefpersoon moet aangeven welke twee figuren uit de twee sets met elkaar overeenkomen).

De adaptatie van de figuurtests aan de computer is beschreven door Maarse en Van de Veerdonk (1991). Bij de conventionele versie van de *Vocabulary* moet de proefpersoon middels het aankruisen van twee woorden zijn of haar antwoord aangeven; in de computer-ondersteunde versie is ervoor gekozen om alle mogelijke paren van woorden (dit zijn er 15) tegelijkertijd aan te bieden. Responsen werden geregistreerd via het toetsenbord. De proefpersonen konden naast de pijltjestoetsen slechts enkele toetsen bedienen: "insert", "home", "page-up", "page-down", "delete" en "end". De andere knoppen waren met een mal afgedekt.

Naast de GATB scores waren voor een deel van de steekproef tevens scores bekend van de *Berenschot Intelligentietest, Figuurseries* (een test waarbij figurale reeksen aangevuld moeten worden), *Instrumentinterpretatie* (een papier-en-potloodtest voor ruimtelijk inzicht; de proefpersoon moet de positie van een vliegtuig aangeven op basis van informatie over de horizon en het kompas) en *Determinationsgerät* (een test voor reactiesnelheid).

De scores op beide GATB versies zijn hier per subtest weergegeven in (een variatie op) zogenaamde "equipercentile equating curves" (zie Figuur 1). Op de horizontale as staan de gestandaardiseerde scores op de conventionele GATB. Op de verticale as is het verschil in gestandaardiseerde score op de conventionele en de computer-ondersteunde test weergegeven. Voor strikt parallelle tests vallen de curves precies op de stippellijn van Figuur 1; voor deze lijn geldt dat de scores gelijk zijn aan nul. Enkele subtests wijken opmerkelijk van de stippellijn af. Om te beginnen *Name Comparison*. De "equipercentile equating curve" van deze test ligt consistent boven de stippellijn. Een positie boven de stippellijn duidt op consistent hogere scores voor de computer-ondersteunde afname. Ook *Tool Matching* laat hogere scores in de computer-ondersteunde afname zien. *Vocabulary* en *Form Matching* laten het omgekeerde patroon zien (consistent hogere scores op de conventionele afnames). De overige tests vertonen geen systematische verschillen tussen beide afnamevormen.


Name Comparison
Computation
Three-Dimensional Space
Vocabulary
Tool Matching
Arithmetic Reasoning
Form Matching
Gestandaardiseerd verschil
Gestandaardiseerde papier-en-potlood score

Figuur 1.

Het aantal items dat per subtest opgelost was, lag hoger voor de computer-ondersteunde versie van de GATB dan voor de conventionele versie. Alleen *Vocabulary* vormde hierop een uitzondering; de conventionele versie gaf hier meer opgeloste items te zien. Het afwijkende patroon van *Vocabulary* is te verklaren vanuit het (eerder beschreven) verschil in stimuluspresentatie; in de conventionele versie leest de proefpersoon zes stimuli en moet zelf een paar vormen, terwijl in de computer-ondersteunde versie de proefpersoon steeds van paren woorden moet nagaan of deze gelijk of tegengesteld zijn. Een belangrijk deel van het verschil in score zit waarschijnlijk dan ook in de tijd die nodig is om de stimuli te lezen.

De proportie correct opgeloste items per subtest was steeds hoger voor de conventionele subtests. Er blijkt sprake te zijn van een interessant verschil in responsestrategie tussen beide testafnames: de computer-ondersteunde afname leidt tot een grotere snelheid en onnauwkeurigheid in het reageren dan de conventionele afname. Het is niet helemaal duidelijk wat de achtergrond van dit verschil in responsestrategie is. Mogelijk heeft het te maken met "demand characteristics" van de testsituatie. Computer-ondersteunde testafnames kunnen nogal eens de suggestie wekken dat er zo snel mogelijk gereageerd moet worden zelfs als de betreffende test niet met een tijdlimiet afgenomen wordt. Computers worden nu eenmaal--om begrijpelijke en door de commercie duidelijk gevoede redenen--met snelheid geassocieerd. Naarmate proefpersonen meer ervaring hebben met computers in het algemeen, zal het verschil in response-strategie tussen de testafnames waarschijnlijk afnemen.

De meta-analyse van Mead en Drasgow (1993) vergeleek de testscores op conventionele en computer-ondersteunde testafnames. In ons eigen onderzoek vonden we dat het aantal correct opgeloste items als score op snelheidstests een onvolledig beeld geeft van de prestatie. Verschillen in de balans tussen snelheid en nauwkeurigheid in beide testafnames worden veronachtzaamd als enkel naar totaalscores gekeken wordt. Naast de itemscores dient ook het aantal opgeloste items geregistreerd te worden.

Vervolgens is een LISREL model gefit om na te gaan of beide tests identieke covariantie-structuren te zien geven. Een dergelijke analyse maakt het mogelijk om verscheidene hypotheses te onderzoeken over de gelijkheid (en verschillen) in beide testversies, zoals de gelijkheid van het aantal factoren, van de ladingen op de factoren en van de correlaties tussen de factoren. Er werd voor beide testmodaliteiten een twee-factormodel gefit. De eerste factor heeft te maken met "perceptual speed" en wordt gedefinieerd door *Name Comparison, Three-Dimenionsal Space* en *Tool Matching* en in iets mindere mate door *Vocabulary*. De tweede factor heeft te maken met rekenvaardigheid en wordt gedefinieerd door *Name Comparison, Computation, Vocabulary* en *Arithmetic Reasoning*. Een model waarin beide de tests in beide versies gelijke ladingen hadden leverde een rede-

lijke fit op. Dit geeft aan dat de psychologische betekenis van de beide testversies nooit ver uiteen kan liggen (ondanks de verschillen in response-strategieën).

Een vergelijkbaar patroon werd ook teruggevonden in de correlaties van de subtests in beide versies met de vier overige taken. De *Berenschot Intelligentietest* vertoonde de hoogste correlaties met de subtests uit beide versies. Dit is niet verwonderlijk omdat zowel de Berenschot als de GATB metingen van intelligentie zijn. *Name Comparison* vertoonde een sterkere correlatie met intelligentie in de computer-ondersteunde versie dan in de conventionele versie. Verder was de relatie van *Form Matching* met de reactiesnelheidstest significant hoger voor de conventionele versie. Het is opvallend dat juist de eenvoudige tests qua betekenis veranderen door deze op de computer te zetten. Het lijkt er sterk op dat vooral "simple, clerical tasks" gevoelig zijn voor verandering in de wijze van afname.

Nieuwe uitdagingen voor het equivalentie-onderzoek

Toen de computer ingevoerd werd om psychologische tests af te nemen, werd al snel de vraag gesteld of de testscores op conventionele en computer-ondersteunde tests wel inwisselbaar waren. Equivalentie-onderzoek begon aanvankelijk met een nauwkeurig geformuleerde vraagstelling maar met weinig hypothesen. Er lijkt de laatste jaren duidelijk sprake te zijn van convergentie van bevindingen; er begint een redelijk consistent beeld te ontstaan. Bij cognitieve taken is met name het snelheidselement van belang. Tests zonder tijdlimiet worden meestal niet of nauwelijks beïnvloed door de introductie van de computer tenzij deze introductie andere wijzigingen met zich meebrengt zoals bij *Vocabulary* in ons eigen onderzoek. Bij persoonlijkheidsvragenlijsten lijkt de invloed van het medium klein als (wederom) maar rekening gehouden wordt met mogelijk onbedoelde verschillen in beide afnamevormen zoals het wegvallen van het persoonlijk contact tussen proefleider en proefpersoon in de computer-ondersteunde testafname.

Als hier beweerd wordt dat de invloed van de introductie van de computer veelal klein is, is daarmee nog niet geïmpliceerd dat de normen van conventionele tests die waarschijnlijk weinig door de introductie van de computer beïnvloed worden (zoals cognitieve taken zonder tijdlimiet waarvan de "look and feel" van beide testversies identiek zijn) ook direct voor computer-ondersteunde afnames bruikbaar zijn. Er doen zich hierbij ten minste twee problemen voor. Om te beginnen hoeven de "equipercentile equating curves" van equivalente tests niet precies op de stippellijn van Figuur 1 te vallen. Alle "equipercentile equating curves" die rechte lijnen zijn, verwijzen naar equivalente tests, maar alleen de tests met curves die op de stippellijn vallen, hebben normen die voor beide testversies identiek zijn. Tests met andere rechte

lijnen als "equating curves" kunnen wel door een simpele lineaire transformatie naar elkaar vertaald worden, maar het zou onjuist zijn identieke normen te gebruiken voor beide tests. Een tweede probleem heeft te maken met de predictieve validiteit. Er is nog steeds niet empirisch aangetoond dat equivalente tests een gelijke predictieve validiteit hebben, ook al is duidelijk dat er tussen equivalente instrumenten geen verschillen te verwachten zijn.

De centrale vraagstelling in equivalentie-onderzoek is in laatste instantie empirisch. Ook al zijn er nog zo weinig redenen om aan te nemen dat de introductie van de computer enige verandering in de aard van de test met zich meebrengt, toch blijft gelden: the proof of the pudding is in the eating. Equivalentie zal steeds empirisch aangetoond moeten worden (Van de Vijver, 1987).

Vanuit theoretisch perspectief blijven nog enkele interessante vragen te beantwoorden in het equivalentie-onderzoek. Deze hebben betrekking op een analyse van de psychologische verschillen tussen beide testmodi. In ons eigen onderzoek vonden we dat met name eenvoudige "snelheids- en nauwkeurigheidstests" beïnvloed werden door de introductie van de computer. Computers lijken personen uit te nodigen om snel maar niet zo nauwkeurig te antwoorden. Het is interessant om na te gaan of de verandering in response-strategie zich door instructies of manipulaties in de testpresentatie laat beïnvloeden. Het is niet ongebruikelijk dat bij een computer-ondersteunde afname na beantwoording van een vraag het volgende item ogenblikkelijk verschijnt. Hiermee kan onwillekeurig en onbedoeld de suggestie opgeroepen worden dat de proefpersoon moet opschieten. Langere intertrial intervallen of een expliciete keuze van de proefpersoon om het volgende item te doen verschijnen ("Druk op een knop om verder te gaan") zouden de snelheid van responderen kunnen veranderen. De achterliggende hypothese zou dan zijn dat het verschil in response-strategie in beide testafnames de verschillen in psychologische betekenis en in nomologisch netwerk kunnen verklaren en dat manipulaties die de verschillen in response-strategie doen verdwijnen ook verschillen in nomologische netwerken elimineren. In dit geval is er sprake van één trek die op twee manieren gemeten wordt. Het is ook mogelijk dat er intrinsieke verschillen bestaan tussen de psychologische processen die opgeroepen worden in beide afnames. In zo'n opvatting zijn de verschillen in onder andere response-strategie een gevolg van het feit dat de vaardigheid of persoonlijkheidseigenschap die gemeten wordt, niet los te zien is van het medium waarin deze gemeten wordt. Snelheid en nauwkeurigheid is dan een vaardigheid die medium-afhankelijk is. Er is dan sprake van twee, mogelijk gecorreleerde trekken die elk middels een eigen medium gemeten worden.

Besluit

Testequivalentie is een belangrijk issue als een testafname voortaan niet meer met papier en potlood maar middels een computer afgenomen wordt. Toch moet het belang van equivalentie voor de praktijk van computer-ondersteunde afnames niet overschat worden. Equivalentie verwijst in principe naar een conservatieve vraagstelling: kan ik in een nieuw medium nog wel de resultaten van het oude medium gebruiken? Maar computer-ondersteunde afnames bieden veel mogelijkheden die met conventionele testafnames niet of moeilijk gerealiseerd konden worden, zoals adaptief testgebruik, de registratie van responsetijden en on-line dataverwerking en -rapportage. De kracht van computer-ondersteunde tests ligt dan ook niet zozeer in hun equivalentie met conventionele tests maar juist in de extra mogelijkheden die gecreëerd worden.

Literatuur

Akkerman, A. E. (1994). Equivalentie als uitdaging! *Psychologie en Computers, 11*, 3-6.

Greaud, V. A., & Green, B. F. (1986). Equivalence of conventional and computer presentation of speed tests. *Applied Psychological Measurement, 10*, 23-34.

Green, B. F. (1991). Guidelines for computer testing. In T. B. Gutkin & J. C. Conoley (Red.), *The computer and the decision making process. Buros-Nebraska symposium on measurement and testing* (pp. 245-273). Hillsdale, NJ: Erlbaum.

Maarse, F. J., & Van de Veerdonk, J. L. A. (1991). Graphical representations in computerized psychological tests. In L. J. M. Mulder, F. J. Maarse, W. P. B. Sjouw, & A. E. Akkerman (Red.), *Computers in psychology. Applications in education, research and psychodiagnostics* (pp. 62-67). Lisse: Swets & Zeitlinger.

Mead, A. L., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin, 114*, 449-458.

Neubauer, A. C., Urban, E., & Malle, B. F. (1991). Raven's Advanced Progressive Matrices: Computerunterstützte Präsentation versus Standardvorgabe. *Diagnostica, 37*, 204-212.

Van de Vijver, F. J. R. (1987). Het gebruik van computer-ondersteunde tests in de diagnostische praktijk. *De Psycholoog, 22*, 10-15.

Van de Vijver, F. J. R., & Harsveld, M. (in druk). The incomplete equivalence of the paper-and-pencil and computerized version of the General Aptitude Test Battery. Journal of Applied Psychology.

Adres auteur:

Fons van de Vijver
Faculteit der Sociale Wetenschappen
Postbus 90153
5000 LE Tilburg